TAMS24: Statistisk teori

• • • Föreläsning 1 • • •

In this lecture, the following definitions are mentioned:

X: random variable (stokastiska variabel);

Mean (Väntevärde):

$$\mu = E(X) = \begin{cases} \sum k p_X(k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_X(x) dx, & \text{if } X \text{ is continuous;} \end{cases}$$

Note:

$$\mu = E(g(X)) = \begin{cases} \sum g(k) p_X(k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx, & \text{if } X \text{ is continuous;} \end{cases}$$

Variance (Varians): $\sigma^2 = V(X) = E((X-\mu)^2) = E(X^2) - (E(X))^2$;

Standard deviation (Standardavvikelse): $\sigma = D(X) = \sqrt{V(X)}$;

There are several properties of mean and variance: X and Y are independent random variables, a,b,c are constants, then

$$\begin{aligned} &E(aX + bY + c) = aE(X) + bE(Y) + c, \\ &V(aX + bY + c) = a^2V(X) + b^2V(Y), \text{ here } X, Y \text{ are independent (oberoende);} \\ &\text{ Note: these two properties also work for } n \text{ random variables.} \end{aligned}$$

If $X \sim N(\mu, \sigma)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$;

If $X_1, \ldots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i)$, then

$$d + \sum_{i=1}^{n} c_i X_i \sim N(d + \sum_{i=1}^{n} c_i \mu_i, \sqrt{\sum_{i=1}^{n} c_i^2 \sigma_i^2});$$

Population X with an unknown parameter θ,

Random sample (slumpmässigt stickprov): $X_1, \ldots, X_n$ are independent and have the same distribution as the population X. Before observe/measure, $X_1, \ldots, X_n$ are random variables.

Observations (observationer): $x_1, \ldots, x_n$ (after observe/measure), which are numbers (not random variables);

Point Estimator (Stickprovsvariabeln): $\hat{\Theta} = f(X_1, \ldots, X_n)$, a random variable;

Point Estimate (Punktskattning): $\hat{\theta} = f(x_1, \ldots, x_n)$, a number;

Unbiased (Väntevärdesriktig): $E(\hat{\Theta}) = \theta$;

Effective (Effektiv): Two point estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased, we say that $\hat{\Theta}_1$ is more effective than $\hat{\Theta}_2$ if $V(\hat{\Theta}_1) < V(\hat{\Theta}_2)$;

Consistent (Konsistent): A point estimator $\hat{\Theta} = g(X_1, \ldots, X_n)$ is consistent if

$$\lim_{n\to\infty} P(|\hat{\Theta} - \theta| > \varepsilon) = 0, \text{ for any constant } \varepsilon > 0.$$

(This is actually called "convergence in probability" in probability and statistics).

Theorem: If $E(\hat{\Theta}) = \theta$ and $\lim_{n\to\infty} V(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is consistent.

TAMS24: Statistisk teori

• • • Föreläsning 2 • • •

Throughout this lecture, a population is denoted as X (with an unknown parameter θ) and a random sample is denoted as $\{X_1, \ldots, X_n\}$ and observations are denoted by $\{x_1, \ldots, x_n\}$.

Commonly used point estimates/estimators
population mean μ: $\hat{\mu} = \bar{x}$ Sample mean (Stickprovsmedelvärde)
Before observe/measure, $\hat{M} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and after observe/measure, $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

population variance σ^2:
(1) If μ is known, Before observe/measure, $\hat{\Sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$,
and after observe/measure, $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$;

(2) If μ is unknown, $\hat{\sigma^2} = s^2$ **Sample variance (Stickprovsvarians):**
Before observe/measure, $\hat{\Sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$,
and after observe/measure, $\hat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n \times \bar{x}^2)$;

Sample standard deviation (Stickprovsstandardavvikelse): Before observe/measure, $S = \sqrt{S^2}$, and after observe/measure, $s = \sqrt{s^2}$;

Method of moments (momentmetoden)—MM: # of equations depends on # of unknown parameters,

$$E(X) = \bar{x},$$

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

$\vdots$

$$E(X^k) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Least square method (minsta-kvadrat-metoden)—LSM: The least square estimate $\hat{\theta}$ is the one minimizing

$$Q(\theta) = \sum_{i=1}^{n}(x_i - E(X))^2.$$

In this lecture, we reviewed several types of random variables:

Binomial distribution $X \sim Bin(N, p)$: there are N independent and identical trials, each trial only has two results: success and failure. Assume the probability of success is p, and $X =$ the number of successes in these N trials. The random variable $X \sim Bin(N, p)$ has a probability mass function (sannolikhetsfunktion)

$$p_X(k) = P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}, k = 1, 2, ..., N;$$

Note: $E(X) = Np$ and $V(X) = Np(1-p)$.

Exponential distribution $X \sim Exp(1/\mu)$: when we consider the waiting time/lifetime... The random variable $X \sim Exp(1/\mu)$ has a density function (täthetsfunktion)

$$f_X(x) = \frac{1}{\mu}e^{-x/\mu}, \quad x \geq 0.$$

Note: $E(X) = \mu$ and $V(X) = \mu^2$.

Poisson distribution $X \sim Po(\mu)$: when we consider number of happenings during the fixed time / length / area / volume. The random variable $X \sim Po(\mu)$ has a probability mass function (sannolikhetsfunktion)

$$p_X(k) = P(X = k) = \frac{\mu^k}{k!}e^{-\mu}, k = 0, 1, 2...;$$

Note: $E(X) = \mu$ and $V(X) = \mu$.

TAMS24: Statistisk teori

• • • Föreläsning 3 • • •

Maximum-likelihood method (Maximum-likelihood-metoden): The maximum-likelihood estimate $\hat{\theta}$ is the one maximizing the likelihood function

$$L(\theta) = \begin{cases} \prod_{i=1}^n f(x_i;\theta), & \text{if } X \text{ is continuous,} \\ \prod_{i=1}^n p(x_i;\theta), & \text{if } X \text{ is discrete.} \end{cases}$$

Remark 1 on ML: In general, it is easier/better to maximize $\ln L(\theta)$;

Remark 2 on ML: If there are several random samples (say m) from different populations with a same unknown parameter θ, then the maximum-likelihood estimate $\hat{\theta}$ is the one maximizing the likelihood function defined as $L(\theta) = L_1(\theta)\dots L_m(\theta)$, where $L_i(\theta)$ is the likelihood function from the i-th population.

Estimates of population variance σ^2: If there is only one population with an **unknown** mean, then method of moments and maximum-likelihood method, in general, give a point estimate of σ^2 as follows

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \qquad \text{(NOT unbiased).}$$

We **adjust/correct** the NOT unbiased point estimate in this way:
We calculate the NOT unbiased point estimator $E(\widehat{\Sigma^2}) = E(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.
To get the unbiased point estimator, that is, to make the expectation equal σ^2, we divide the coefficient $\frac{n-1}{n}$, we get the new point estimator $\widehat{\Sigma^2} = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \times \sum_{i=1}^n (X_i - \bar{X})^2$.
You can check the new point estimator $E(\frac{1}{n-1} \times \sum_{i=1}^n (X_i - \bar{X})^2) = \sigma^2$, which is unbiased. So

an **adjusted (or corrected)** point estimate would be the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 \qquad \text{(unbiased).}$$

If there are m different populations with unknown means and a same variance σ^2, then an **adjusted (or corrected)** ML estimate is

$$s^2 = \frac{(n_1-1)s_1^2 + \ldots + (n_m-1)s_m^2}{(n_1-1) + \ldots + (n_m-1)} \qquad \text{(unbiased)}$$

where n_i is the sample size of the i-th population, and s_i^2 is the sample variance of the i-th population.

Standard error (medelfelet) of an point estimate $\hat{\theta}$=an estimation of $D(\hat{\Theta})$= an estimation of $\sqrt{V(\hat{\Theta})}$;

TAMS24: Statistisk teori

• • • Föreläsning 4 • • •

In this lecture, we talked about two new types of random variables: $t(f)$***-fördelning*** and $\chi^2(f)$***-fördelning***. The exact definitions of these random variables are not important. We focused on the graphs of these random variables and found various critical values in the following forms, for instance,

$$\lambda_{0.025} = 1.96, \qquad t_{0.025}(30) = 2.04, \qquad \chi^2_{0.025}(30) = 47, \qquad \chi^2_{0.975}(30) = 16.8, \qquad ......$$

Throughout this lecture, we have a random sample $\{X_1, \dots, X_n\}$ from $N(\mu, \sigma)$.

1.1 $(1-\alpha)$ **confidence interval (konfidensintervall)** I_μ **for** μ (by the way $(1-\alpha)$ is called confindence coefficient (*konfidensgrad*))
(a). If σ is known, then the fact is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, and therefore

$$I_\mu = \bar{x} \mp \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

(b). If σ is unknown, then the fact is $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, and therefore

$$I_\mu = \bar{x} \mp t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}.$$

1.2. $(1-\alpha)$ **confidence interval (konfidensintervall)** I_{σ^2} **for** σ^2 (or I_σ for σ)
The fact is $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, and therefore

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \qquad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \right)$$

Remark. All intervals in above **1** and **2** are two-sided (tvåsidigt). In the lecture, we also worked on several intervals which are one-sided (ensidigt) in the forms $(-\infty, b)$ and $(a, +\infty)$.

TAMS24: Statistisk teori

• • • Föreläsning 5 • • •

In this lecture, we had three topics:

(1) confidence intervals for two (or more) random samples from normal distributions.

One sample $\{X_1, \dots, X_n\}$ from $N(\mu, \sigma)$:

$$I_\mu = \begin{cases} \bar{x} \mp \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \text{ if } \sigma \text{ is known; } \left[\text{the fact } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\right] \\ \bar{x} \mp t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \text{ if } \sigma \text{ is unknown; } \left[\text{the fact } \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)\right] \end{cases}$$

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right); \left[\text{the fact } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)\right]$$

Unknown σ^2 can be estimated by the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$

Two samples $\{X_1, \dots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$, $\{Y_1, \dots, Y_{n_2}\}$ from $N(\mu_2, \sigma_2)$, $N(\mu_1, \sigma_1)$ indep. $N(\mu_2, \sigma_2)$:

$$I_{\mu_1-\mu_2} = \begin{cases} (\bar{x}-\bar{y}) \mp \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \text{ if } \sigma_1 \text{ and } \sigma_2 \text{ are known;} \\ \qquad \left[\text{the fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)\right] \\ (\bar{x}-\bar{y}) \mp t_{\alpha/2}(n_1+n_2-2)\cdot s \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \text{ if } \sigma_1 = \sigma_2 = \sigma \text{ is unknown;} \\ \qquad \left[\text{the fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)\right] \\ \approx (\bar{x}-\bar{y}) \mp t_{\alpha/2}(f)\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \text{ if } \sigma_1 \neq \sigma_2 \text{ both are unknown;} \\ \qquad \left[\text{the fact } \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \approx t(f)\right] \\ \qquad \left[\text{degrees of freedom } f = \frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}\right] \end{cases}$$

$$I_{\sigma^2} = \left(\frac{(n_1+n_2-2)s^2}{\chi^2_{\frac{\alpha}{2}}(n_1+n_2-2)}, \quad \frac{(n_1+n_2-2)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n_1+n_2-2)}\right), \text{ if } \sigma_1 = \sigma_2 = \sigma;$$

$$\left[\text{the fact } \frac{(n_1+n_2-2)S^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)\right]$$

Unknown σ^2 can be estimated by the samples variance $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

m samples: The unknown $\sigma_1^2 = \dots = \sigma_m^2 = \sigma^2$ can be estimated by $s^2 = \frac{(n_1-1)s_1^2+\dots+(n_m-1)s_m^2}{(n_1-1)+\dots+(n_m-1)}$.

An important example: The idea of using hjälpvariabel to find confidence intervals is EXTREMELY important. There are a lot more different confidence intervals besides above. For instance, we consider two independent samples: $\{X_1, \dots, X_{n_1}\}$ from $N(\mu_1, \sigma)$ and $\{Y_1, \dots, Y_{n_2}\}$ from $N(\mu_2, \sigma)$. In this case, we can easily prove that

$$c_1\bar{X} + c_2\bar{Y} \sim N\left(c_1\mu_1 + c_2\mu_2, \quad \sigma\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}\right).$$

- If σ is known, then the fact $\frac{(c_1\bar{X}+c_2\bar{Y})-(c_1\mu_1+c_2\mu_2)}{\sigma\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}} \sim N(0,1)$. So we can find $I_{c_1\mu_1+c_2\mu_2}$;
- If σ is unknown, then the fact $\frac{(c_1\bar{X}+c_2\bar{Y})-(c_1\mu_1+c_2\mu_2)}{S\sqrt{\frac{c_1^2}{n_1}+\frac{c_2^2}{n_2}}} \sim t(n_1+n_2-2)$. So we can find $I_{c_1\mu_1+c_2\mu_2}$.

Questions to think about: In the above example, what if we have two populations $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$ with $\sigma_1 \neq \sigma_2$? (two cases: both σ_1 and σ_2 are known; both σ_1 and σ_2 are unknown).

(2) confidence intervals from normal approximations.

$$X \sim Bin(N,p):\ I_p = \hat{p} \mp \lambda_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}},\ \text{the fact } \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{N}}} \approx N(0,1).$$

(we require that $N\hat{p} > 10$ and $N(1-\hat{p}) > 10$)

$$X \sim Hyp(N,n,p):\ I_p = \hat{p} \mp \lambda_{\alpha/2}\sqrt{\frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\hat{p}(1-\hat{p})},\ \text{the fact } \frac{\hat{P}-p}{\sqrt{\frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\hat{P}(1-\hat{P})}} \approx N(0,1).$$

$$X \sim Po(\mu):\ I_\mu = \bar{x} \mp \lambda_{\alpha/2}\sqrt{\frac{\bar{x}}{n}},\ \text{the fact } \frac{\bar{X}-\mu}{\sqrt{\frac{\bar{X}}{n}}} \approx N(0,1).$$

(we require that $n\bar{x} > 15$)

$$X \sim Exp(\frac{1}{\mu}):\ \bullet\ I_\mu = \left(\frac{\bar{x}}{1+\frac{\lambda_{\alpha/2}}{\sqrt{n}}}, \frac{\bar{x}}{1-\frac{\lambda_{\alpha/2}}{\sqrt{n}}}\right),\ \text{the fact } \frac{\bar{X}-\mu}{\mu/\sqrt{n}} \approx N(0,1),$$

$$\bullet\ I_\mu = \bar{x} \mp \lambda_{\alpha/2}\frac{\bar{x}}{\sqrt{n}},\ \text{the fact } \frac{\bar{X}-\mu}{\bar{X}/\sqrt{n}} \approx N(0,1).$$

(we require that $n \geq 30$)

An important example: Again, the use of the fact to find confidence intervals is EXTREMELY important. There are more confidence intervals besides above. For instance, we consider two independent samples: X from $Bin(N_1,p_1)$ and Y from $Bin(N_2,p_2)$, with unknown p_1 and p_2. As we know that

$$\hat{P}_1 \approx N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \text{ and } \hat{P}_2 \approx N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right),$$

so $\hat{P}_1 - \hat{P}_2 \approx N\left(p_1-p_2, \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}\right)$. Therefore, the fact is $\frac{(\hat{P}_1-\hat{P}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \approx N(0,1)$,

$$I_{p_1-p_2} = (\hat{p}_1-\hat{p}_2) \mp \lambda_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

(4) Large sample size ($n \geq 30$, population may be completely unknown).

If there is no information about the population(s), then we can apply Central Limit Theorem (usually with a large sample $n \geq 30$) to get an approximated normal distributions. Here are two examples:

Example 1: Let $\{X_1, \ldots, X_n\}, n \geq 30$, be a random sample from a population, then (no matter what distribution the population is)

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0, 1).$$

Example 2: Let $\{X_1, \ldots, X_{n_1}\}, n_1 \geq 30$, be a random sample from a population, and $\{Y_1, \ldots, Y_{n_2}\}, n_2 \geq 30$, be a random sample from another population which is independent from the first population, then (no matter what distributions the populations are)

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0, 1).$$

Final remark of this lecture: Ideally, you should be able to derive/prove all these confidence intervals after this lecture. I strongly suggest you at least try to prove all these. It is VERY important that you understand all (for instance, you should feel easy to derive all the corresponding *one-sided* confidence intervals).

TAMS24: Statistisk teori

• • • Föreläsning 6 • • •

A new topic: ***Hypothesis testing*** (***Hypotesprövning***).

In this lecture, we focused on ***Hypothesis testing*** without Normal (aprroximation) and the *general theory* of hypothesis testing. Namely, there is a random sample $\{X_1, \dots, X_n\}$ from a population X with an unknown parameter θ,

$$H_0: \ \theta = \theta_0 \qquad vs. \qquad H_1: \ \theta < \theta_0, \text{ or } \theta > \theta_0, \text{ or } \theta \neq \theta_0$$

	H_0 is true	H_0 is false and $\theta = \theta_1$
reject H_0	(type I error or significance level) α	(power) $h(\theta_1)$
don't reject H_0	$1 - \alpha$	(type II error) $\beta(\theta_1) = 1 - h(\theta_1)$

We also talked about *p-value* and mentioned that

$$\underline{\text{reject } H_0} \textit{ if and only if } \underline{p\text{-value} < \alpha}.$$

In computer lab 1, you will use the confidence intervals from the ratio of two population variances.

In order to study this, we need a new distribution F***-fördelning***: If $X \sim \chi^2(r_1)$ is independent of $Y \sim \chi^2(r_2)$, then $\frac{X/r_1}{Y/r_2} \sim F(r_1, r_2)$. (here r_1 and r_2 are degrees of freedom)

Now suppose we have two independent samples $\{X_1, \dots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$, and $\{Y_1, \dots, Y_{n_2}\}$ from $N(\mu_2, \sigma_2)$. We have already known that $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$ and $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$, so by definition

$$\text{the fact} \qquad \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Therefore

$$I_{\sigma_2^2/\sigma_1^2} = \left(\frac{s_2^2}{s_1^2} \cdot F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), \quad \frac{s_2^2}{s_1^2} \cdot F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \right).$$

TAMS24: Statistisk teori

• • • Föreläsning 7 • • •

Continuation of hypothesis testing: We considered special cases of hypothesis testing using a test statistic directly related to the parameter of interest. Compare the test statistic with the fact in confidence intervals, and try to understand the equivalence between *hypothesis testing* and *confidence intervals*!!! Throughout the lectures,

$$\begin{aligned}
&\text{TS} := \text{“test statistic”} \;---\; \text{depends on the fact and } H_0\\
&\text{C} := \text{“rejection region ”=“critical region”} \;---\; \text{depends on the fact and } H_1
\end{aligned}$$

$\underline{\text{reject } H_0}$ *if* $\underline{\text{TS} \in \text{C}}$;

$\underline{\text{reject } H_0}$ *if and only if* $\underline{p\text{-value} < \alpha}$.

(1) Hypothesis testing for population mean(s).

One sample: $\{X_1, \ldots, X_n\}$ from $N(\mu, \sigma)$. Null hypothesis $H_0 : \mu = \mu_0$.

$$\begin{cases}
\begin{array}{c}\sigma \text{ is known:}\\ \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\end{array}
\begin{cases}
H_1 : \mu < \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}},\ \text{C} = (-\infty, -\lambda_\alpha),\\
\qquad p\text{-value} = P(N(0,1) \le \text{TS});\\
H_1 : \mu > \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}},\ \text{C} = (\lambda_\alpha, +\infty),\\
\qquad p\text{-value} = P(N(0,1) \ge \text{TS});\\
H_1 : \mu \ne \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}},\ \text{C} = (-\infty, -\lambda_{\alpha/2}) \cup (\lambda_{\alpha/2}, +\infty),\\
\qquad p\text{-value} = 2P(N(0,1) \ge |\text{TS}|).
\end{cases}\\[2ex]
\begin{array}{c}\sigma \text{ is unknown:}\\ \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)\end{array}
\begin{cases}
H_1 : \mu < \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (-\infty, -t_\alpha(n-1)),\\
\qquad p\text{-value} = P(t(n-1) \le \text{TS});\\
H_1 : \mu > \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (t_\alpha(n-1), +\infty),\\
\qquad p\text{-value} = P(t(n-1) \ge \text{TS});\\
H_1 : \mu \ne \mu_0 : \ \text{TS} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}},\ \text{C} = (-\infty, -t_{\alpha/2}(n-1)) \cup (t_{\alpha/2}(n-1), +\infty),\\
\qquad p\text{-value} = 2P(t(n-1) \ge |\text{TS}|).
\end{cases}
\end{cases}$$

Two samples: $\{X_1, \ldots, X_{n_1}\}$ from $N(\mu_1, \sigma_1)$; $\{Y_1, \ldots, Y_{n_1}\}$ from $N(\mu_2, \sigma_2)$; Null hypothesis $H_0 : \mu_1 = \mu_2$.

$$
\begin{cases}
\begin{array}{c}\sigma_1, \sigma_2 \text{ are known:} \\ \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)\end{array}
\begin{cases}
H_1 : \mu_1 < \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (-\infty, -\lambda_\alpha), \\
\qquad p\text{-value} = P(N(0,1) \leq \text{TS}); \\
H_1 : \mu_1 > \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (\lambda_\alpha, +\infty), \\
\qquad p\text{-value} = P(N(0,1) \geq \text{TS}); \\
H_1 : \mu_1 \neq \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \ \text{C} = (-\infty, -\lambda_{\alpha/2}) \ \cup \ (\lambda_{\alpha/2}, +\infty), \\
\qquad p\text{-value} = 2P(N(0,1) \geq |\text{TS}|).
\end{cases} \\
\begin{array}{c}\sigma_1 = \sigma_2 \text{ is unknown:} \\ \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)\end{array}
\begin{cases}
H_1 : \mu_1 < \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (-\infty, -t_\alpha(n_1+n_2-2)), \\
\qquad p\text{-value} = P(t(n_1+n_2-2) \leq \text{TS}); \\
H_1 : \mu_1 > \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (t_\alpha(n_1+n_2-2), +\infty), \\
\qquad p\text{-value} = P(t(n_1+n_2-2) \geq \text{TS}); \\
H_1 : \mu_1 \neq \mu_2 : \ \text{TS} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \ \text{C} = (-\infty, -t_{\alpha/2}(n_1+n_2-2)) \\
\qquad \qquad \cup \ (t_{\alpha/2}(n_1+n_2-2), +\infty), \\
\qquad p\text{-value} = 2P(t(n_1+n_2-2) \geq |\text{TS}|).
\end{cases} \\
\sigma_1 \neq \sigma_2 \text{ both unknown: similarly as in the tree of confidence intervals.}
\end{cases}
$$

(2) Hypothesis testing for population variance(s).

$$
\begin{cases}
\begin{array}{c}\{X_1, \ldots, X_{n_1}\} \text{ from } N(\mu, \sigma) \\ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \\ H_0 : \sigma^2 = \sigma_0^2\end{array}
\begin{cases}
H_1 : \sigma^2 < \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2}, \ \text{C} = (0, \chi^2_{1-\alpha}(n-1)), \\
\qquad p\text{-value} = P(\chi^2(n-1) \leq \text{TS}); \\
H_1 : \sigma^2 > \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2}, \ \text{C} = (\chi^2_\alpha(n-1), +\infty), \\
\qquad p\text{-value} = P(\chi^2(n-1) \geq \text{TS}); \\
H_1 : \sigma^2 \neq \sigma_0^2 : \ \text{TS} = \frac{(n-1)s^2}{\sigma_0^2}, \ \text{C} = (0, \chi^2_{1-\frac{\alpha}{2}}(n-1)) \ \cup \ (\chi^2_{\frac{\alpha}{2}}(n-1), +\infty), \\
\qquad p\text{-value} = 2P(\chi^2(n-1) \geq \text{TS}) \text{ or } 2P(\chi^2(n-1) \leq \text{TS}).
\end{cases} \\
\begin{array}{c}\{X_1, \ldots, X_{n_1}\} \text{ from } N(\mu_1, \sigma_1) \\ \{Y_1, \ldots, Y_{n_2}\} \text{ from } N(\mu_2, \sigma_2) \\ \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1) \\ H_0 : \sigma_1^2 = \sigma_2^2\end{array}
\begin{cases}
H_1 : \sigma_1^2 < \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2, \ \text{C} = (0, F_{1-\alpha}(n_1-1, n_2-1)), \\
\qquad p\text{-value} = P(F(n_1-1, n_2-1) \leq \text{TS}); \\
H_1 : \sigma_1^2 > \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2, \ \text{C} = (F_\alpha(n_1-1, n_2-1), +\infty), \\
\qquad p\text{-value} = P(F(n_1-1, n_2-1) \geq \text{TS}); \\
H_1 : \sigma_1^2 \neq \sigma_2^2 : \ \text{TS} = s_1^2/s_2^2, \ \text{C} = (0, F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)) \\
\qquad \qquad \cup \ (F_{\frac{\alpha}{2}}(n_1-1, n_2-1), +\infty), \\
\qquad p\text{-value} = 2P(F(n_1-1, n_2-1) \geq \text{TS}) \\
\qquad \qquad \text{or } 2P(F(n_1-1, n_2-1) \leq \text{TS}).
\end{cases}
\end{cases}
$$

(3) Large sample size ($n \geq 30$, population may be completely unknown): If there is no information about the population(s), then we can apply Central Limit Theorem (usually with a large sample $n \geq 30$). The idea is exactly the same as the one used in confidence intervals. **One example** is: a sample

$\{X_1, \ldots, X_n\}, n \geq 30$, from some population (which is unknown) with a mean μ and standard deviation σ. Null hypothesis $H_0 : \mu = \mu_0$. Then it follows from CLT that $\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1)$, therefore

$$\begin{cases} H_1 : \mu < \mu_0 : & \text{TS } = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C } = (-\infty, -\lambda_\alpha), \\ & p\text{-value} = P(N(0,1) \leq \text{TS}); \\ H_1 : \mu > \mu_0 : & \text{TS } = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C } = (\lambda_\alpha, +\infty), \\ & p\text{-value} = P(N(0,1) \geq \text{TS}); \\ H_1 : \mu \neq \mu_0 : & \text{TS } = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}, \ \text{C } = (-\infty, -\lambda_{\alpha/2}) \ \cup \ (\lambda_{\alpha/2}, +\infty), \\ & p\text{-value} = 2P(N(0,1) \geq |\text{TS}|). \end{cases}$$

Besides confidence intervals, we briefly mentioned *Prediktionsintervall*. Roughly speaking, a prediktionsintervall is an interval for a newly selected element, while a confidence interval is for some unknown parameter (mean or variance), not for a specific element.

TAMS24: Statistisk teori

• • • Föreläsning 8 • • •

We have a NEW topic in this lecture: *Multi-dimension random variables* (or *random vectors*), which are related to *linear regressions.*

Covariance (Kovarians) of (X,Y)**:** $\sigma_{X,Y} = cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$,
$cov(X,X) = V(X)$ and $cov(X,Y) = cov(Y,X)$.

Correlation coefficient (Korrelation) of (X,Y)**:** $\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{V(X)\cdot V(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}$.

A rule: for real constants a, a_i, b and b_j,

$$cov(a + \sum_{i=1}^{m} a_i X_i, b + \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j cov(X_i, Y_j).$$

X **and** Y **are uncorrelated:** if $\rho(X,Y) = 0$, i.e. $cov(X,Y) = 0$.

An important theorem: Suppose that a random vector $\mathbf{X}$ has a mean $\mu_{\mathbf{X}}$ and a covariance matrix $C_{\mathbf{X}}$. Define a new random vector $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, for some matrix A and vector $\mathbf{b}$. Then

$$\mu_{\mathbf{Y}} = A\mu_{\mathbf{X}} + \mathbf{b}, \quad C_{\mathbf{Y}} = AC_{\mathbf{X}}A'.$$

Standard normal vectors: $\{X_i\}$ are independent and $X_i \sim N(0,1)$,

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \text{ thus } \quad \mu_{\mathbf{X}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad C_{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \text{ density } f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{x}'\mathbf{x}}.$$

General normal vectors: $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where $\mathbf{X}$ is a standard normal vector, and

$$\mu_{\mathbf{Y}} = \mathbf{b}, \quad C_{\mathbf{Y}} = AA', \quad \text{density } f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(C_{\mathbf{Y}})}} e^{-\frac{1}{2}\left[(\mathbf{y}-\mu_{\mathbf{y}})' C_{\mathbf{Y}}^{-1} (\mathbf{y}-\mu_{\mathbf{y}})\right]}.$$

Independent and Uncorrelated:
1. If X and Y are independent, then X and Y are uncorrelated. Conversely, generally, if X and Y are uncorrelated, we can't say X and Y are independent. For example:
If we have the following random variables X, Y,

X	0	1
$p_X(k)$	0.5	0.5

and

Y	-1	1
$p_X(k)$	0.5	0.5

Then we can get

XY	-1	0	1
$p_{XY}(k)$	0.25	0.5	0.25

and

X^2Y	-1	0	1
$p_{X^2Y}(k)$	0.25	0.5	0.25

Now we let $Z = XY$, we can see X and Z are not independent!!!
But $cov(X,Z) = E(XZ) - E()XE(Z) = 0$, that is X and Z are uncorrelated!!!
2. **If** X **and** Y **are jointly normally distributed, then Uncorrelated implies independent.**

TAMS24: Statistisk teori

• • • Föreläsning 9 • • •

Simple and Multiple linear regressions (Enkel och Multipel linjär regression) are the main topic.

Simple linear regression: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_j \sim N(0,\sigma), i = 1, \ldots, n.$

Multiple linear regression: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma), i = 1, \ldots, n.$

Both 'Simple linear regression' and 'Multiple linear regression' can be written as vector forms:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}: \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n}).$$

$$\mathbf{Y} \sim N(\mu_{\mathbf{Y}}, C_{\mathbf{Y}}), \text{ where } \mu_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} \text{ and } C_{\mathbf{Y}} = \sigma^2 \mathbf{I}_{n\times n}.$$

Estimate of the coefficient $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$.

Estimator of the coefficient $\boldsymbol{\beta}$: $\hat{\boldsymbol{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$.

Estimated regression line is: $\hat{\mu} = y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$.

Analysis of variance:

$$SS_{TOT} = \sum_{j=1}^{n} (y_j - \bar{y})^2, \quad \frac{SS_{TOT}}{\sigma^2} = \frac{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}{\sigma^2} \sim \chi^2(n-1), \text{ if } \beta_1 = \ldots = \beta_k = 0;$$

$$SS_R = \sum_{j=1}^{n} (\hat{\mu}_j - \bar{y})^2, \quad \frac{SS_R}{\sigma^2} = \frac{\sum_{j=1}^{n} (\hat{\mu}_j - \bar{Y})^2}{\sigma^2} \sim \chi^2(k), \text{ if } \beta_1 = \ldots = \beta_k = 0;$$

$$SS_E = \sum_{j=1}^{n} (y_j - \hat{\mu}_j)^2, \quad \frac{SS_E}{\sigma^2} = \frac{\sum_{j=1}^{n} (Y_j - \hat{\mu}_j)^2}{\sigma^2} \sim \chi^2(n-k-1).$$

$$SS_{TOT} = SS_R + SS_E, \text{ and } R^2 = \frac{SS_R}{SS_{TOT}}.$$

$* * *$ σ^2 is estimated as $\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1}$.

$* * *$ For the Hypothesis testing: $H_0 : \beta_1 = \ldots = \beta_k = 0$ vs $H_1 :$ at least one $\beta_j \neq 0$,

$$\begin{cases} \frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1) \\ \text{TS} = \frac{SS_R/k}{SS_E/(n-k-1)} \\ C = (F_\alpha(k, n-k-1), +\infty). \end{cases}$$

$* * *$ We know $\hat{\boldsymbol{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$, thus if we denote

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{pmatrix} h_{00} & h_{01} & \cdots & h_{0k} \\ h_{10} & h_{11} & \cdots & h_{1k} \\ \vdots & \vdots & & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kk} \end{pmatrix},$$

then $\hat{B}_j \sim N(\beta_j, \sigma\sqrt{h_{jj}})$ and $\frac{\hat{B}_j - \beta_j}{\sigma\sqrt{h_{jj}}} \sim N(0,1)$. But σ is generally unknown, therefore

$$\frac{\hat{B}_j - \beta_j}{S\sqrt{h_{jj}}} \sim t(n-k-1), \qquad \left[s\sqrt{h_{jj}} \text{ is sometimes denoted as } d(\hat{\beta}_j) \text{ or } se(\hat{\beta}_j)\right].$$

Confidence interval of β_j is: $I_{\beta_j} = \hat{\beta}_j \mp t_{\alpha/2}(n-k-1) \cdot s\sqrt{h_{jj}}$;

Hypothesis testing $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ has

$$\begin{cases} \text{TS} = \frac{\hat{\beta}_j}{s\sqrt{h_{jj}}} \\ C = (-\infty, -t_{\alpha/2}(n-k-1)) \ \cup \ (t_{\alpha/2}(n-k-1), +\infty). \end{cases}$$

TAMS24: Statistisk teori

• • • Föreläsning 10 • • •

Continued: Simple and Multiple linear regressions (Enkel och Multipel linjär regression):

$$\begin{aligned}
&Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma), \quad \text{(the model)};\\
&\mu = E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \quad \text{(the mean)};\\
&\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \quad \text{(the estimated regression line)}.
\end{aligned}$$

For a given/fixed $\mathbf{x} = (1, x_1, \ldots, x_k)'$, $\hat{\mu}$ is an estimate of unknown μ (and Y). Then we can talk about 'accuracy' of this estimate in terms of confidence intervals (and prediction intervals).

Confidence interval of μ: $I_\mu = \hat{\mu} \pm t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{\mathbf{x}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}}$.

Prediction interval of Y: $I_Y = \hat{\mu} \pm t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{1 + \mathbf{x}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}}$.

Suppose we have two models:

$$\begin{cases}
\text{Model 1:} \quad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon;\\
\text{Model 2:} \quad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k+p} x_{k+p} + \varepsilon,
\end{cases}$$

and we want to test $H_0 : \beta_{k+1} = \ldots = \beta_{k+p} = 0$ vs $H_1 :$ at least one $\beta_{k+i} \neq 0$,

$$\begin{cases}
\frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1)\\
\text{TS} = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)}\\
C = (F_\alpha(p, n-k-p-1), +\infty).
\end{cases}$$

TAMS24: Statistisk teori

• • • Föreläsning 11 • • •

For χ^2**-test**, we have two parts. The first part is about χ^2-test of pupulation:

$$\begin{aligned} H_0 :& X \sim \text{ distribution (with or without unknown parameters)};\\ H_1 :& X \nsim \text{ distribution} \end{aligned}$$

Then

$$\begin{aligned} \text{fact } :& \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \sim \chi^2(k - 1 - \#\text{of unknown parameters});\\ \text{TS} =& \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i};\\ C =& \left(\chi^2_\alpha(k - 1 - \#\text{of unknown parameters}),\ \ +\infty\right). \end{aligned}$$

The second part is about χ^2-test of *Homogeneity* (independence). Suppose we have a data with r rows and k columns,

$$\begin{aligned} H_0 :& \text{ grouping in rows and grouping in columns are independent (i.e. they don't affect each other)};\\ H_1 :& \text{ grouping in rows and grouping in columns are NOT independent (i.e. they affect each other))}. \end{aligned}$$

Then

$$\begin{aligned} \text{fact } :& \sum_{j=1}^{k}\sum_{i=1}^{r} \frac{(N_{ij} - np_{ij})^2}{n\mathrm{p}_{ij}} \sim \chi^2((r-1)(k-1));\\ \text{TS} =& \sum_{j=1}^{k}\sum_{i=1}^{r} \frac{(N_{ij} - np_{ij})^2}{n\mathrm{p}_{ij}};\\ C =& \left(\chi^2_\alpha((r-1)(k-1)),\ \ +\infty\right). \end{aligned}$$